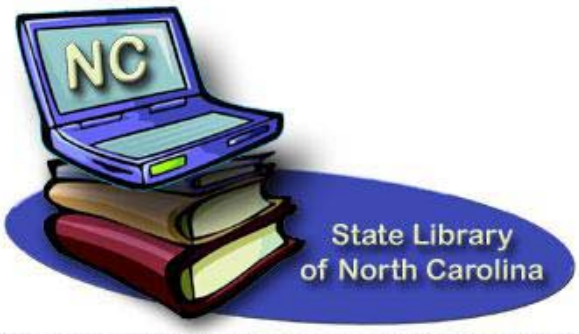

Access to State Government Information

Solutions Work Group Preservation Committee

Meeting: February 20, 2004

Present:	Druscie Simpson	Paolo Mangiafico
	Kelly Eubank	Kristin Martin
	Helen Tibbo	Jan Reagan
	Bob Brinson	Lucy Reid
	Karrie Peterson	

The Preservation Committee discussed the current processes for collecting, managing, and making paper information accessible, for both the State Library and State Archives. Following that, the Committee discussed research into preserving digital information and critiqued the proposed process to create a digital repository. While the State Archives and Records needs were considered here as, the process focused on the State Library's roles in the proposed digital archive. The Committee then discussed potential digital repository systems and provided recommendations for next steps.

Current Document Collection Systems:

The process for collecting state publications by the State Library and the process for collecting records by Archives and Records was discussed.

State Library Process: North Carolina State Publications Clearinghouse (Condensed from handout at the meeting)

1. Collects ten printed state publications from all state agencies within ten days of issuance. Each agency has a publications officer responsible for sending publications to the Library.
2. Processes the publications, produces microfiche copies of all publications for distribution, catalogs all publications, and submits MARC records to OCLC and the State Library's Voyager catalog. Two copies are added to the permanent State Library collection.
    a. Serials and monographs are cataloged, while ephemera is not.
3. Distributes remaining paper copies and microfiche copies to depository library around the state.

State Archives and Records (In addition to discussing the current process, some issues relating to records changing format, such as digital imaging, were discussed in the context of the current process. Thanks to Druscie and Kelly who provided this information)

1. The custodial agency and records analyst from the North Carolina State Archives determine the proper retention and disposition of all public records created by that agency. The head of the custodial agency and the administrative staff of the North Carolina. State Archives agree upon this disposition.
    a. Schedules may not be updated as frequently as records series change.
2. The custodial agency transfers the records in question at its scheduled time by calling the State Records Center and ordering containers, a transfer form, and custom labels.
    a. The process is reactive, not proactive. The Government Records Branch relies on agencies to contact it at the appropriate time.
3. Once the records are boxed and properly labeled, staff from the State Records Center get the records and bring them to one of the three storage areas of the Government Records Branch in the Archives and Records Section.
4. Once transferred, the records remain in the State Records Center until the end of a prescribed retention period. They then are destroyed or transferred to the State Archives.
    a. While records are stored in the Records Center, the agency retains legal custody of the records.
    b. The index of the records transferred to the Records Center is currently created and maintained by the agency. It probably should be scheduled as well, so that the archives can have a copy for processing.
5. If the paper records are permanently valuable and transferred to the State Archives, an archivist is assigned to appraise, arrange, and describe the records so that easy access to the records is provided.
    a. Approximately 15% of all records are transferred to the archives. From there, the amount of records permanently stored is further winnowed down to about 5%.
    b. Records transferred to Archives may encounter a processing backlog.
6. The clients are predominantly state agencies and select county agencies. The Administrative Office of the Courts (AOC) manages county government records created after court reform. Registers of Deeds transfer some of their records to the Archives, but they also have series of records that they maintain in the county, such as Deed Books. Other clients are universities and community colleges but they generally have their own archives.

High Level Models for Digital Repositories

The committee discussed both the Library of Congress's proposed digital repository model and the Open Archival Information System Reference Model. More information on the Library of Congress's proposed architecture can be found in Appendix 9 of *Preserving Our Digital Heritage: Plan for the National Digital Information Infrastructure and Preservation Program*. Washington, D.C.: Library of Congress, 2002. http://www.digitalpreservation.gov/index.php?nav=3&subnav=1. A new article introducing OAIS for the layperson, describes OAIS in more detail: Lavoie, Brian F. "The Open Archival Information System Reference Model: Introductory Guide." In *DPC Technology Watch*. 04-01. Dublin, Ohio: OCLC, 2004. http://www.dpconline.org/graphics/reports/index.html#intoais. (Thanks for Helen for finding the article). Druscie recently attending a meeting of the Southern

Archives and Records Conference. Other states have been grappling with similar issues of digital preservation. North Carolina is the only state in the south that has not combined responsibility for digital records and publications into a single agency. States are at different levels in creating their own digital repositories, but most are looking at the issue too.

Tiers of Custodianship (Levels of Digital Management)
Paolo described how Duke University is looking at supporting the long-term preservation of digital information and the three levels of support they are providing. Expenses increase with each higher level.

1. Stored – Capture now, deal with later. Basic indexing, but limited access and metadata about the information stored.
2. Preserved – information is stored in native format, but has basic metadata describing it, and is more easily accessible.
3. Archived – full selection process, normalized supported formats for long-term preservation, complete metadata.

Proposed Digital Repository
The State Library and Archives and Records have created a vision for a proposed digital repository for capturing and storing digital state information. A draft of the process and steps toward creating a digital repository from the State Library perspective is provided below. The State Archives and Records has many additional issues and this process would form part of a larger system that would deal with digital information of all types, including records that are not publications or websites. Since the Preservation Committee meeting, the State Library and Archives and Records have worked together to create a unified vision of a digital repository. However, the Preservation Committee discussion focused on issues and concerns specific to the Library process. While many concerns were raised during the discussion, the basic fundamental process and steps for the repository creation remain unchanged.

1. **Content Input/Collection**
    a. All state agency information (content) published on the web must go through the records management application (RMA).
    b. Metadata (created using the NC GILS standard, subset of NC GILS, or other standard such as Dublin Core) must accompany all content. Ideally, metadata should be generated automatically to keep the process quick and simple for the content providers; however, in some cases, content providers will have to manually create metadata.
        i. Manually created metadata (created by content creators) might include:
            1. Descriptive information, e.g. Title, Originator, possibly subject terms
            2. Identification of type of work
                a. traditional publication
                b. scheduled record
                c. webpage

3. Level of web page as described by Web Content Analysis and a level of change status (new, minor changes, major changes, complete overhaul)

ii. Automatically generated metadata (created by system) might include:

1. Linkage (different temporal versions of work; parent-child relationships)
2. Rights information (copyright, etc.)
3. Date of upload
4. Preservation metadata about computer programs (should be detected automatically)
5. Structural metadata to bind together different components of a webpage (e.g. identifying images in webpage, different file formats combined together, different webpages that should be linked together into one object)

c. Older webpages/digital publications and content for agencies that choose not to participate with the RMA will be collected using a web-crawling tool.
d. The State Library will select and appraise all content – deposited through the RMA or collected through the crawl - as outlined in Step 3 below.

2. **Content Assignment and Metadata Storage**
   a. Content should be deposited in "triage" where it will be automatically sorted and
      i. published on the web ("live version")
      ii. distributed out to appropriate virtual communities (e.g. traditional publications, records, webpages) stored within the digital repository according to its metadata ("digital repository version")
   b. The repository version should be assigned a URI for identification.
   c. Metadata should be stored separately from the content (object?)
   d. Metadata should point to the both versions of the content:
      i. Digital repository version
      ii. Live version available on the web

3. **Digital Clearinghouse (Library processing)**
   a. Clearinghouse staff will select and appraise incoming content before permanent repository assignments are made. This selection and appraisal should be semi-automated.
      i. Legacy collection publications will be printed out, cataloged, and stored permanently in the State Library Documents Collection.
      ii. Metadata for traditional publications identified as serials and monographs will be exported to the State Library catalog (Voyager) for more detailed description in MARC format and EnCompass, another part of the Library's ILS that provides a gateway for access to information wherever it lives. The digital repository must be accessible via the Library's ENCompass system, (i.e., links to digital repositories, other library catalogs, etc.)

iii. The Library might supplement the descriptive metadata accompanying ephemera and webpages that would remain in the digital repository to improve searching (e.g. with subject terms)
iv. The Library will delete all content it deems inappropriate or unnecessary for the digital repository. (As the library fine-tunes its selection criteria, content (objects?) deemed inappropriate or unnecessary should be automatically blocked from storage in the repository).

4. **The Digital Repository**
    a. The repository should be searchable and available to the public unless agencies specifically request a certain work not be available (rights restrictions or some other reason).
        i. The search should search through full-text and metadata of the repository as well as full-text of the "live" web.
        ii. Searches may be limited to only the repository or only the web.
    b. Content for permanent preservation should be easily exported from the repository, along with its corresponding metadata, and transferred to a "preservation repository" in the future. Content for permanent preservation will be selected according to criteria developed by the State Library. In general, traditional-style publications will have a higher priority than ephemera and webpages.

Issues and Concerns with the Digital Repository Process:

1. If every published to the web must be submitted through the Records Management Application, compliance issues would be solved, but the repository could be inundated with too much information.
2. The creation of metadata is problematic. We would like the creator to provide at least some descriptive metadata, which will aid us in identifying and classifying the information, as well as deciding whether the digital object should be selected for permanent retention. However, there is a limit as to how much metadata creators would be willing to create, and the quality of metadata could be lower than that done by an information professional. As much metadata as possible should be automatically generated.
3. It is difficult to categorize information. We will need good definitions to be able to define whether it is a publication or a webpage.
4. Records that aren't produced via the web would not be captured. We would need a separate system for capturing them.
5. ASP (Active Server Pages) and other dynamically generated webpages can be difficult to capture and may change too frequently for us to get every iteration.
6. We will need to identify the level of change to a website and determine the level of risk based on the Web Content Assessment levels to figure out which objects should be stored in the digital repository. The producer would probably have to determine this, or possibly the web server could measure the level of change. (Web Content Assessment table is with the Identification/Selection Committee meeting notes).
7. We will need to have a way to triage the objects entering the repository to determine the level of preservation and access support.

8. The system of capture only works for newly produced information. We will need to be able harvest information from existing websites or from websites where agencies choose not to participate.
9. We'll need a way to verify authenticity of the information in the repository.
10. We need to consider how to save the objects: content vs. context. Some publications might be better pulled from their website, while other times an individual might want to see the publications in its original environment of the surrounding website.
11. Webpages cross the boundary between traditional publications and traditional records. The State Library and Archives and Records will need to work together to determine responsibilities and methods of access.
12. Any digital repository would need to have support from statewide ITS and the support of the agencies involved in creating materials.
13. We will need to have incentives for the state agencies to be involved. The following are some ideas:
    a. create ease in web publishing
    b. make information easier to archive
    c. improve workflow
    d. agencies can comply with library and records using one system
    e. introduce awards or threaten with audit reports for compliance
    f. have different levels of participation for state agencies, so they can choose what works best for them and provide monetary awards for agencies that fully participate in the repository program, e.g. create their metadata
    g. have agency-level IT approve all web publishing
14. Metadata created for the repository could be made available for general harvesting via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).
15. Any system used for a digital repository will need to have open standards to ensure long-term viability.

Potential Digital Repository Systems

1. Documentum

Documentum is a Records Management Application (RMA). ITS has invested resources to purchase licenses for 100 seats of two silos of the Documentum – Enterprise Document Management and Web Content Management. The Library and the State Archives and Records would be interested in having additional functionality added for records management from the Fixed Content Management System and the capability to deal with rich media through one more silo. ITS is interested in working with the Library and Archives and Records to create a digital repository using Documentum. We are considering it because:

1. It's already been purchased.
2. The State of Michigan and the San Diego Supercomputer Center have been using the software to test long-term preservation issues.
3. It addresses possible incentives for state agencies by helping them manage their document work flow and have single system for compliance with archives and the Library.
4. Has been tested by Michigan before and has worked out some of the kinks.
5. It would allow capture of web documents at the point of creation.

DSpace
DSpace is an institutional repository system, in use by some universities. For more information on the background of DSpace, see http://www.dspace.org. Both Duke and North Carolina State University libraries are testing DSpace currently. It is open-source. Duke chose it for testing because it is one of the most popular repository programs being tested, is relatively user friendly, flexible, and has heavy user support. The search mechanism is rather simple. UNC is also testing DSpace with a project collecting materials from retiring professors. We also discussed Fedora, an open-source digital repository system created by Cornell University and the University Virginia, but have concerns that although it has more sophisticated technology, it won't be supported long-term. The OCLC Digital Archive could be another option for a digital repository and is something we should explore. There is some concern about expense and the long-term viability of relying on a private company like OCLC.

Summary of Challenges and Issues
The committee concluded by summarizing six challenges of digital preservation

1. Capturing content – ingest.
2. When do we create metadata and how much do we create?
3. Identifying and organizing a huge amount of material and ensuring authenticity of the information.
4. Determining where content falls into the tiers of custodianship.
5. How to handle long-term preservation, e.g. renderability, functional preservation versus preserving the whole experience (content vs. context).
6. Financial sustainability (while not specifically mentioned at the meeting, Karrie brought this point up in an email).

Preservation Committee Recommendations

1. Pilot project using Documentum as a digital repository. We will be attending a demonstration on March 15 to learn more. The Library will work with Archives and Records on this project.
2. While not discussed in detail by the committee, project staff recently learned that we will be able to test software developed by the Illinois State Library designed to crawl and capture state government websites. We can use the captured material to test other digital repository systems.
3. Investigate the feasibility of using DSpace for storing state government information. There may be options for collaboration with area universities. First we could download our own version and play around with the functionality. We'll be capturing websites through a project sponsored by the Illinois State Library, and can use these to test the functionality of DSpace.
4. Investigate storing documents on the ENCompass server, at least as a temporary measure. We can use websites captured through the Illinois State Library Project to test functionality.
5. Find out more information about the OCLC digital archive. Talk to original pilot members of the archive, talk to OCLC about possibilities.